Estimating Public Opinion with Large Language Models

Agustina Pesce[1], Adrian Ruhe[1], Sofía Karsaclian[1]

[1]University of Konstanz

agustina.pesce@uni-konstanz.de, adrian.ruhe@uni-konstanz.de, sofia.karsaclian@uni-konstanz.de

Abstract

Survey research plays a crucial role in informing policy decisions, yet challenges such as non-response rates and attrition have risen in the past decades. Large Language Models (LLMs) offer a promising cost-effective solution to provide in-silico answers that can serve as results screenings to assist survey polls. However, implementing LLMs for social sensing presents challenges, including biased behavior and limited context understanding. In this study, we evaluate the quality of LLMs in predicting survey outcomes on public opinion matters using data from the American National Election Studies (ANES) 2020 survey. We also investigate how LLM performance is influenced by exposure to social media discussions on specific topics, using Reddit comments. Our analysis reveals mixed evidence regarding the accuracy of LLM predictions, with variations across different prompts and models. We find that while LLMs can provide reasonably accurate predictions for some policy issues, they may also exhibit biases and produce polarized outcomes, particularly when exposed to social media discussions. We discuss the implications of these findings and highlight the importance of considering platform demographics and discussion context in LLM-based social sensing.

1. Introduction

Survey research has provided valuable input for policymakers to make decisions on economic strategy and public health (Chu et al., 2023). In the last decades, despite the money and time invested in surveys (Jansen et al., 2023), they have shown increasingly difficult problems regarding their non-response rate and attrition, affecting results accuracy and representativeness (Bisbee et al., 2023; Chu et al., 2023).

Due to their practical consequences and rising problems, finding solutions to the usual pitfalls of surveys with automated pipelines can be of high interest to researchers and governments. Large Language Models (LLMs) have been investigated as a possible answer for some of these shortcomings thanks to their capabilities to mimic humanlike answers, low costs and fast performance (Jansen et al., 2023). However, the new problems that arise with them still make their best implementation unsure, being until now implemented with different methodological approaches.

In this project, we test the quality of LLMs results when prompted to create in-silico social sensing on public opinion matters. We do this using data from the ANES 2020 survey on specific public opinion issues. Furthermore, we want to understand how the model's performance changes when exposed to social media discussions on specific topics. For this, we use Reddit's comments on submissions with news articles on the selected topics.

In-silico surveys with LLMs

Previous research on LLM implementations for survey results prediction has leveraged the promising capabilities LLMs have to adopt personas. By stating different sociodemographic group belongings, the LLMs create tailored answers towards public opinion or group affinity questions.

Bisbee et al. (2023), for example, created synthetic personas for ChatGPT 3.5 Turbo and asked it to rate a series of groups using thermometer scores as in the American National Election Survey (ANES). In their results, they observed that average synthetic opinions were often substantively and statistically indistinguishable. However, the answers altogether showed a bias towards greater affective polarization by showing greater in-group preference and out-group rejection. Additionally, the in-silico answers were more homogenous, accounting for a smaller standard deviation, which could not be compensated with higher model temperatures.

Open source (OS) models have also been used with the aim of creating synthetic survey responses. Compared to black-box models, OS models offer the opportunity to be further trained for specific tasks by changing their parameters through fine-tuning. In the research by Chu et al. (2023), a BERT model was fine-tuned with the publications of four major mass media outlets. The fine-tuned model had a greater performance at predicting individual's attitudes towards COVID-19. However, in the field of consumer confidence, questions regarding personal matters such as individuals' financial situation or housing value had only low or negative correlations with their ground truth counterparts. As explained by the authors, these results align with findings that show that news coverage mostly affects sociocentric and prospective attributes.

Implementing LLMs for questions on social matters carries risks related to their biased behavior due to their training conditions. To perform next token prediction and

learn insights on correct language generations, LLMs are trained on unsupervised tasks with a large amount of text. Given that social groups do not have equal access to collaborating in these datasets, also called "documentation debt", models are trained with a bias towards hegemonic viewpoints (Bender et al., 2021). In the area of in silico surveys, Besbee et al. (2023) results showed more accurate outcomes for the opinions of Non-Hispanic Whites than those of non-Hispanic black and Hispanic. In addition, other survey prediction implementations have shown models' lower accuracy for individuals with low socioeconomic status (Kim & Lee, 2023). Therefore, using these technologies as replacement of surveys filled out by individuals may disrupt decisions willing to be made on democratic groundings and should be avoided.

Context learning and context extension

One limitation of LLMs is their limited context-window given that they are trained on fixed length sequences. Recently, a large effort has been put into developing and improving LLM capabilities by increasing the context length. This enables the users to provide more information to the model via prompts.

One way of increasing the context-length is via fine-tuning. One example is CodeLlama, a fine-tuned version of Llama 2 trained on sequences of 16.000 tokens, which provides stable generations with up to 100.000 tokens of context (Rozière et al., 2024).

Many other extension techniques are derivations of Rotary Position Embeddings (RoPE), a method that leverages position embedding values that vary along a predictable smooth rotation though the different token positions (Su et al., 2023). This positional embedding behavior allows for a better performance for prompts that exceed the training length. Some examples are Positional Interpolation (Chen et al., 2023), YaRN (Peng et al., 2023), Resonance RoPE (Wang et al., 2023) and PoSE (Zhu et al., 2024).

These methodologies assume that LLMs lack the inherent capability to understand long contents, and typically also require fine tuning for the extension. Conversely, a novel approach called SelfExtend (Jin et al., 2024) proposes that LLMs already have the capabilities to handle long context even with shorter pre-training context windows. SelfExtend expands existing LLMs context windows by constructing bi-level attention information -grouped attention and neighbor attention- computed based on the original model's self-attention mechanism during inference. Without any fine-tuning, SelfExtend efficiently extends the context window up to 24,000 tokens maintaining model performance. In addition, this method can be complemented with flash attention and quantization to reduce the computation power needed for the long prompt processing.

Social sensing with LLMs

Accounting for this background, in this project we propose to analyze the accuracy with which closed source (GPT-3.5) and open source models can predict the outcome of a survey on social matters from different areas of interest. In order to achieve this, we simplify previous approaches by prompting the models to output the frequency distribution of categorical questions.

In other words, instead of creating in-silico answers for the different personas, the aim of this research is to assess the model's ability to perform social sensing. We utilize the 2020 American National Election Studies (ANES) Time Series Study as a benchmark to evaluate discrepancies in the frequency distributions generated by the models. Our objectives include quantifying the extent of these variations and assessing whether the model exhibits any inclination towards specific political attitudes. Our selected questions consist of some of the most regarded topics from the year of the study (2020): immigration, abortion, gun control, climate change and public health (Pew Research Center, 2020).

In addition, this project also aims to leverage the capability of LLMs to be sensitive to in-context learning. To achieve this, models will be inputted at prompt time with discussions over the topics of interest from the platform Reddit. Our aim is to evaluate whether LLMs can adjust their initial estimations based on the provided discussion context. In essence, we seek to determine whether the "social sensing" capabilities of LLMs can be calibrated toward a specific representative group, potentially mitigating biases inherent in the training data.

Using users' discussions will lessen the shortcomings of mass media data, by directly collecting citizens' opinions on the matter. In that sense we seek to find out whether a LLM can be adjusted so that it effectively mirrors public opinion and yields results closer to real public opinion surveys, and therefore being able to replace costly polls and surveys. Moreover, this adaptive capability isn't restricted to public opinion but can extend to many other areas.

To input Reddit's discussions to the model, we decided to explore the SelfExtend method, which we believe can provide a more efficient alternative to fine-tuning.

3. Data

American National Election Studies (ANES)

ANES is a collaboration of Duke University, University of Michigan, The University of Texas at Austin, and Stanford University. The study is conducted every 4 years among American voters, both before and after presidential elections. Respondents are queried about various demographic factors before being asked for their opinions on a wide array of political and societal issues. Our selected topics align with those anticipated to be most significant in the 2020 presidential election (Pew Center Research, 2020). Furthermore, we believe our chosen topics immigration, -abortion, gun control, climate change and public health- are

known for evoking drastically divergent opinions and therefore also likely to be controversially discussed on social media.

Reddit

In order to perform in-context learning for our models, we need suitable data that preferably reflects a wide range of opinions on our chosen topics. We decided to get the data from Reddit via the PRAW API, which provides easy and efficient retrieval of submissions and comments. All discussions were held in the r/news subreddit, a subreddit with 28 million members, where news articles and current events of the United States are discussed. Given that it does not explicitly align with any particular party or ideology, this subreddit provides the most representative forum for extracting discussions.

Nevertheless, as with most social media platforms, Reddit's demographics differ from what a representative sample of the country's population would be. According to the platform, 58% of the users are aged between 18 and 34 years old, and 57% of the users are males (Reddit, 2021). For the 2016 election, Barthel et al. (2016) stated that only 4% of US adults used Reddit. Furthermore, Reddit is a very international platform, so even in US-related subreddit there is a participation of non US citizens. This is a limitation, but also an opportunity to explore how these different demographics affect the LLMs' estimations.

Using the API, we obtained a preselection of submissions from r/news dating from 2020 related to the questions selected from the ANES survey, which also surpassed a determined number of comments. Afterwards, we manually selected the final submissions based on its linked news article; the main requisites were that it was US-based, and that it consisted solely (or as solely as possible) on the topic in question.

4. Methods

Models

To implement the in-context learning approach to survey social sensing, two LLMs were used. First, we implemented the state-of-the-art closed model "GPT-3.5-Turbo-0125" from OpenAI. This model has a context window of up to 16,385 word-piece tokens which allows the input of all prompt tokens without any extensions.

Secondly, we used "Mistral-7B-Instruct-v0.1", a generative text model from Mistral AI. Although the model works with a sliding window that allows for greater prompt extensions than those 4.096 with which it was trained (Jiang et al., 2023), it is not able to process the content of longer prompts as a whole. To address this limitation, we employed the above mentioned SelfExtend, to be able to input a substantial amount of reddit discussion data. Jin et al. provide an example to demonstrate the effectiveness of their method, which we also implemented in our code. While the basic mistral model is only able to filter out a passkey hidden in an otherwise nonsensical text if the token length of the does not exceed its sliding window length, SelfExtend makes it possible to find the passkey in a text of up to 32,000 tokens. The method uses a custom-made flash attention and quantization to enhance performance while maintaining low computational requirements. Nevertheless, running it required the use of the largest GPU available on Google Colab (A100).

We also tried to implement CodeLlama from Meta AI, which we could successfully run with context windows of up to 8,000 tokens. However, the model refused to answer when faced against the prompts, stating that the question was not real but hypothetical and therefore, not possible to answer.

Prompts: social and Reddit sensing

Two versions of social sensing prompts were inputted into the model. Firstly, models were exposed to prompts with the social sensing instruction, the number of survey subjects, the corresponding year of the ANES survey and the ANES question with its corresponding answer categories and requested to estimate the frequency distribution (zero-shot). An example of this prompt for the abortion issue is presented below:

> 'In the following you will be given a question followed by several possible answer categories. Please estimate how a representative sample of *7900* US-Americans would have answered the question in 2020 by providing the number of respondents for each answer option in percent (output a dictionary object with answer category number as key and answer frequency in percent as value. Only provide the number without the percentage sign). Here is the question: *There has been some discussion about abortion during recent years. Which one of the opinions on this page best agrees with your view? 1. By law, abortion should never be permitted 2. The law should permit abortion only in case of rape, incest, or when the woman's life is in danger 3. The law should permit abortion other than for rape/incest/danger to woman but only after need clearly established 4. By law, a woman should always be able to obtain an abortion as a matter of personal choice*'.

Secondly, for the social sensing objective, both models were also run with prompt versions that included 100 Reddit comments on submissions related to the question's topic from 2020, in order to further learn opinion specificities of the time.

Finally, to have a more accurate measure of Reddit discussion leanings compared to ANES survey responses, models were tasked with "strictly" deriving their answers from the provided comments.

The prompts were carried out for five runs in each of the models for every question. The responses were then

averaged across the runs to provide representative answers from the models.

While the results given by "GPT-3.5-Turbo-0125" were mostly satisfying and corresponded to the instructions given, the answers from "Mistral-7B-Instruct-v0.1" exhibited several structural and content-related flaws. This is why the answers from the latter model required multiple preprocessing steps. Issues included the model always filling every answer with the maximum available tokens, resulting in many unwanted characters or wrong dictionary structures and data types. However, the main problem was that Mistral performs poorly with basic mathematical operations. That is why we removed the number of respondents from the prompt, hoping to obtain reasonable percentage-based answers. As this was not the case, we normalized the results to 100 to mimic the relationship between the answers given. While many responses fell within the range of 90 to 110, suggesting a reasonable normalized frequency distribution, some answers did not sum up properly, and occasionally, the model assigned identical percentages to every answer, resulting in invalid outcomes, as indicated by the following results.

5. Results

The models' outcomes and the ANES results were visualized in bar plots to illustrate the difference between the models' outputs and their corresponding ground truth values (Figure 1, Appendix Figure 3). In this visualization, no clear hierarchy for the accuracy performance of the prompting methods (only instruction, with comments, with strict comments) can be observed.

Boxplots with the answers distributions were created to display the outputs' means and standard deviations (Figure 2, Appendix Figure 4). These boxplots correspond to bootstrapped values (1000 iterations) for the distributions' means and standard deviations to get more robust comparison measures through this re-sampling technique. For the GPT model, while the comments improved results for the topics of abortion and climate change, they took to worse predictions for the gun control, immigration and healthcare questions. In the case of Mistral, comments improved the abortion, climate change and immigration answers. In regard to the boxplots dispersion, it can be mentioned that it visibly increases for the strict comments prompt in the GPT model. Also here, we cannot observe an evident trend depending on the model used. While for some questions mean and standard deviation are very close to the ground truth, others differ more strongly.

As summary measures, the average differences and the average absolute difference were calculated (Table 1). For the GPT model, the instruction prompt without any further information was the best performing to predict the ANES survey results with an outstanding average difference of 0.04. However, when absolute differences were accounted for, the comments prompt was the most accurate (0.45). For Mistral, on the other hand, the best average difference corresponded to the strict version prompt for the comments and the best absolute average difference to the comments alone prompt type.

Finally, as suggested by previous research, the average standard deviation ratios between the LLM and the ground truth survey were computed (Table 1). Unlike the LLM implementations with personas, our results do not show smaller standard deviations than the original survey. In this regard, the Mistral model with instruction prompt was the model with the closest dispersion as compared to the ground truth distribution (1% greater) followed by the GPT model with the instruction prompt (4% greater). Although the average dispersions closely align with the ground truth answer, there may be more significant deviations for individual cases. It is also noticeable that the input of comments results in a considerably greater dispersion. Especially when the model is strictly based on the comments, we can see a distinct increase in peripheral opinions which can also be seen in the boxplots.

To sum up, we can say that while both LLMs provide reasonable results on average, it's apparent that their distributions of answers for specific questions can diverge clearly from those of the ANES survey. While the GPT model tends to exhibit greater variability in its responses based on which prompts are used, the results for the Mistral model are more uniformly distributed (as it might just follow a given distribution). Moreover, the Mistral model occasionally produces uniform answer frequencies across all possibilities, making the outcome unreliable.

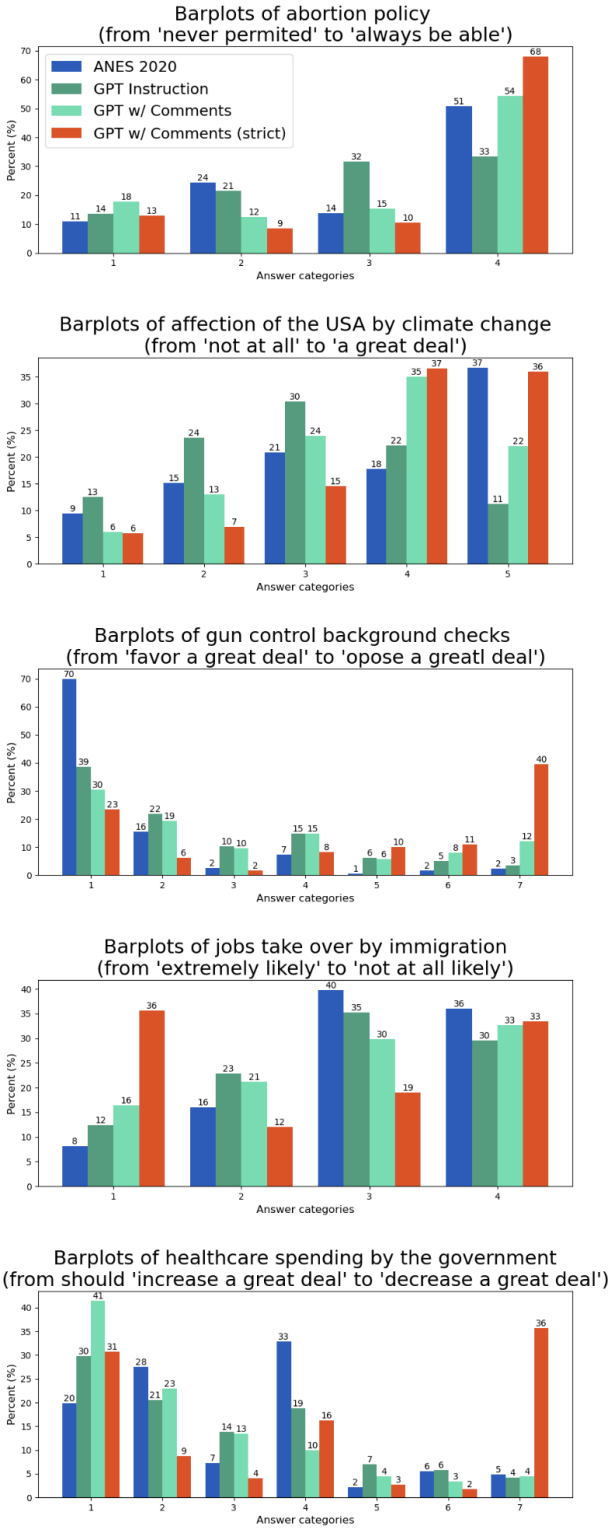


Figure 1: Barplots of ANES (2020) percentages and GPT-3.5-Turbo-0125 social and Reddit sensing.

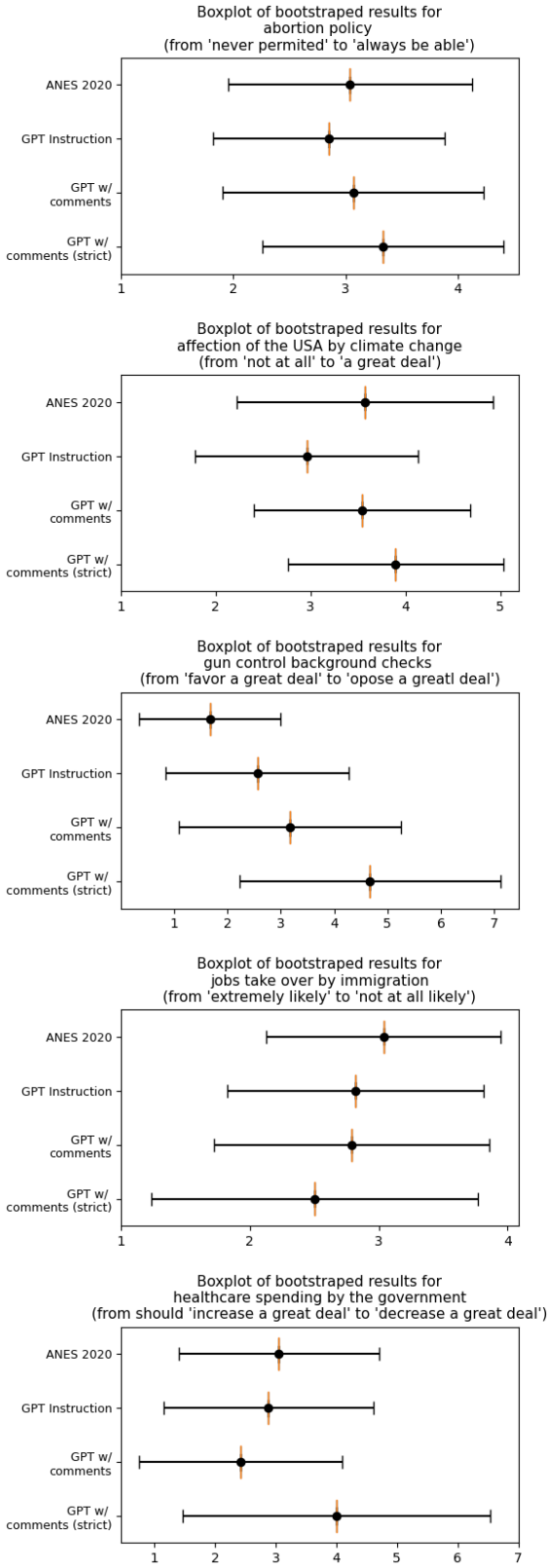


Figure 2: Boxplots of ANES (2020) and GPT-3.5-Turbo-0125 frequency distributions for social and Reddit sensing.

	Average mean diff.	Average mean diff. (abs)	Std. dev. ratio (LLM/ANES)
GPT Instruction	0.04	0.48	*1.04*
GPT w/ comments	-0.31	0.45	1.16
GPT w/ comments (strict)	-0.77	1.03	1.26
Mistral Instruction	0.34	0.97	1.01
Mistral w/ comments	-0.30	*0.83*	1.10
Mistral w/ comments (strict)	*-0.22*	0.95	1.21

Table 1: Std. dev. ratios and average differences for all survey questions for all models and prompts tested. Note: overall best results are underlined and the models' best results are in italics.

6. Discussion and conclusions

In this project, the effectiveness of LLM's social sensing capabilities through zero-shot in-context learning were tested. According to our results, the GPT-3.5-Turbo-0125 model has greater skills to output accurate survey predictions as compared to open-source Mistral-7B-Instruct-v0.1 in the context of policy opinions of USA based respondents.

The comparisons of the different prompt types within models showed mixed evidence. While comments or even the strict use of comments sometimes enhanced the accuracy of the models predicting the opinions on some policy issues, others were worsened by them. This outcome underlines the relevance of choosing a discussion platform that portrays the population's opinion representatively and, in the case of Reddit, the importance of the specific subreddit or submission.

Regarding the results' dispersion, on average for all methods used, the LLMs produced results with a greater standard deviation which could be interpreted as more polarized answer distributions. This tendency was further increased when the models were exposed to Reddit comments and even more when they were asked to base their answers strictly on them. These results suggest that when LLM models are implemented for policy social sensing, an overrepresentation of the population polarization may occur, a problem that seems intensified by social media discussion data.

Several limitations of the performed analysis must be addressed. In first place, when performing social sensing while working with data from social media platforms one has to be aware that the platforms' user demographics differ from what a representative sample of the countries' population. We recognize Reddit's limitations in this sense.

Secondly, the selection of comments for the prompts was restricted to just one submission per topic. As mentioned before, the submissions in r/news correspond to a news article. Therefore, although we sought to obtain articles as closely related to the topic (with no other interfering topics), this is of course not possible in real life. It would be interesting to explore larger discussion datasets and other public opinion sources, such as parliamentary debates, news articles, or other social media platforms.

Another potential concern is that the models may have been trained on the survey results from 2020, so they would not be estimating but rather reproducing the data. Because of the black-box nature of OpenAI, it is not possible to check whether or not they are trained on ANES data. Although possible for some Open Source models, this was also not possible for Mistral, since they claim they cannot share details about the training and datasets due to the highly competitive nature of the field.

Finally, and as mentioned in the results section, the Mistral model showed severe limitations when faced with the task of estimating distributions. Context extension methods like SelfExtend are very useful to increase the model's understanding of larger inputs (as seen in the passkey example) but can unfortunately not account for the models weakness in mathematical operations.

In sum, although LLM's social sensing capabilities could not show accurate results for all policies in this study, the positive effect of in-context learning provided by social media discussions for some of the topics encourages further developments in this research direction. Novel investigations with a focus on using more powerful models or finding discussions with good representative qualities may make this cost-effective and fast approach a moderately reliable social sensing instrument which can serve as a tool to complement survey polls.

7. References

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *Proceedings of the 40th International Conference on Machine Learning, 202*.

American National Election Studies. (2004) *Home.* ANES | American National Election Studies. Retrieved March 24, 2024, from https://electionstudies.org/

Barthel, M., Stocking, G., Holcomb, J., & Mitchell, A.. (2016, February 25). Seven-in-Ten Reddit Users Get News on the Site. *Pew Research Center's Journalism Project.* https://www.pewresearch.org/journalism/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/

Bisbee, J., Clinton, J. D., Dorff, C., & Larson, J. (2023). *Artificially Precise Extremism: How Internet-Trained LLMs Exaggerate Our Differences*. https://doi.org/10.31235/osf.io/5ecfa

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2023). *Synthetic Replacements for Human Survey Data? The Perils of Large Language Models* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/5ecfa

Chen, S., Wong, S., Chen, L., & Tian, Y. (2023). *Extending Context Window of Large Language Models via Positional Interpolation* (arXiv:2306.15595). arXiv. https://doi.org/10.48550/arXiv.2306.15595

Chu, E., Andreas, J., Ansolabehere, S., & Roy, D. (2023). *Language Models Trained on Media Diets Can Predict Public Opinion* (arXiv:2303.16779). arXiv. http://arxiv.org/abs/2303.16779

Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2024). *Questioning the Survey Responses of Large Language Models* (arXiv:2306.07951). arXiv. http://arxiv.org/abs/2306.07951

Garcia, D., Pellert, M., Lasser, J., & Metzler, H. (2021). *Social media emotion macroscopes reflect emotional experiences in society at large* (arXiv:2107.13236). arXiv. http://arxiv.org/abs/2107.13236

Jansen, B. J., Jung, S., & Salminen, J. (2023). Employing large language models in survey research. *Natural Language Processing Journal*, *4*, 100020. https://doi.org/10.1016/j.nlp.2023.100020

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B* (arXiv:2310.06825). arXiv. https://doi.org/10.48550/arXiv.2310.06825

Jin, H., Han, X., Yang, J., Jiang, Z., Liu, Z., Chang, C.-Y., Chen, H., & Hu, X. (2024). *LLM Maybe LongLM: Self-Extend LLM Context Window Without Tuning* (arXiv:2401.01325). arXiv. http://arxiv.org/abs/2401.01325

Kim, J., & Lee, B. (2023). *AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction* (arXiv:2305.09620). arXiv. http://arxiv.org/abs/2305.09620

Peng, B., Quesnelle, J., Fan, H., & Shippole, E. (2023). *YaRN: Efficient Context Window Extension of Large Language Models* (arXiv:2309.00071). arXiv. https://doi.org/10.48550/arXiv.2309.00071

Pew Research Center. (2020, August 13). Important issues in the 2020 election. *Pew Research Center - U.S. Politics & Policy*. https://www.pewresearch.org/politics/2020/08/13/important-issues-in-the-2020-election/

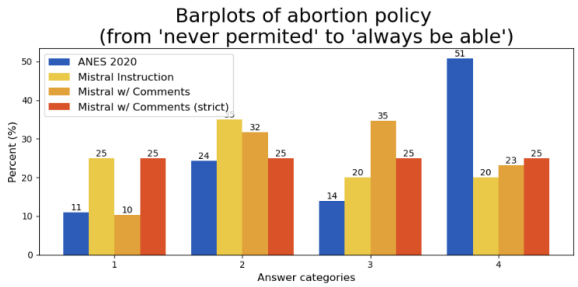
Reddit. (2021, January 17). *Advertising—Audience—Reddit*. https://web.archive.org/web/20210117184818/https://www.redditinc.com/advertising/audience

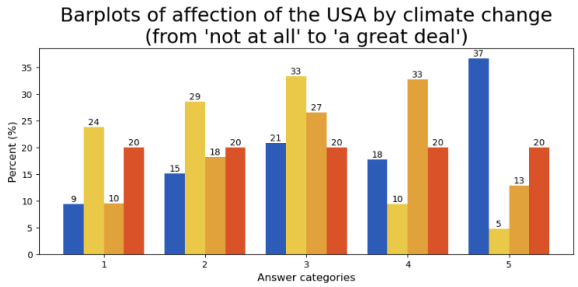
Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., Rapin, J., Kozhevnikov, A., Evtimov, I., Bitton, J., Bhatt, M., Ferrer, C. C., Grattafiori, A., Xiong, W., Défossez, A., … Synnaeve, G. (2024). *Code Llama: Open Foundation Models for Code* (arXiv:2308.12950). arXiv. https://doi.org/10.48550/arXiv.2308.12950

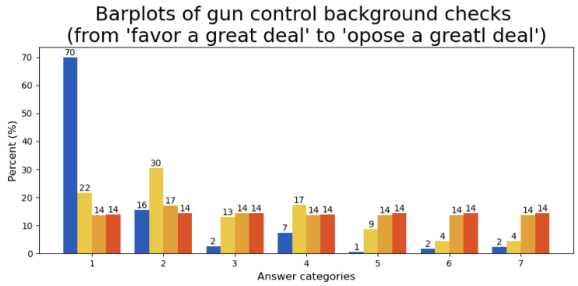
Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2023). *RoFormer: Enhanced Transformer with Rotary Position Embedding* (arXiv:2104.09864). arXiv. https://doi.org/10.48550/arXiv.2104.09864

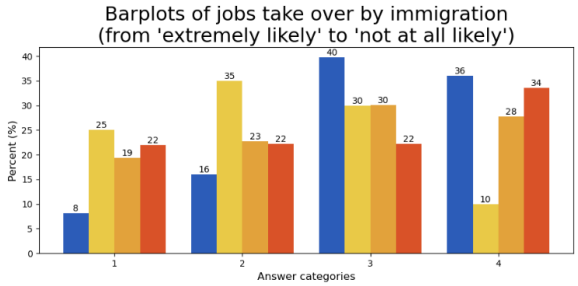
Wang, S., Kobyzev, I., Lu, P., Rezagholizadeh, M., & Liu, B. (2024). *Resonance RoPE: Improving Context Length Generalization of Large Language Models* (arXiv:2403.00071). arXiv. https://doi.org/10.48550/arXiv.2403.00071

Zhu, D., Yang, N., Wang, L., Song, Y., Wu, W., Wei, F., & Li, S. (2024). *PoSE: Efficient Context Window Extension of LLMs via Positional Skip-wise Training* (arXiv:2309.10400). arXiv. https://doi.org/10.48550/arXiv.2309.10400









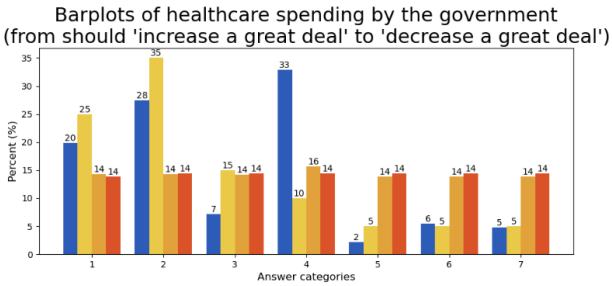
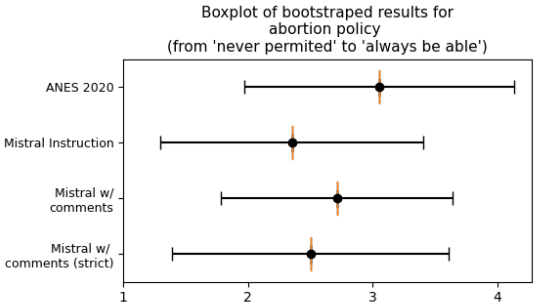
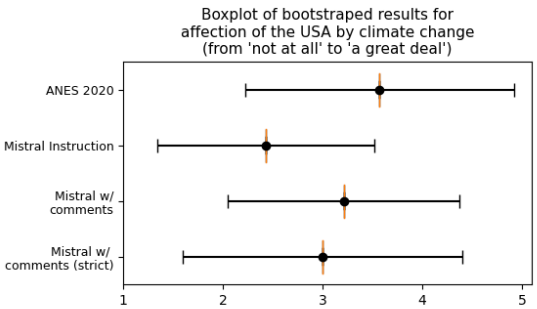
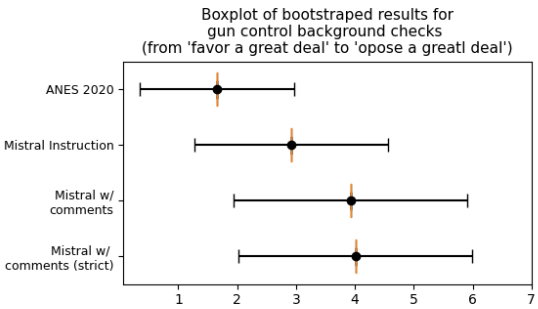
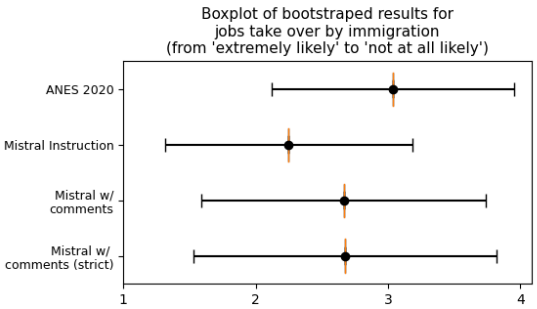


Figure 3: Barplots of ANES (2020) percentages and Mistral-7B-Instruct-v0.1 social and Reddit sensing.









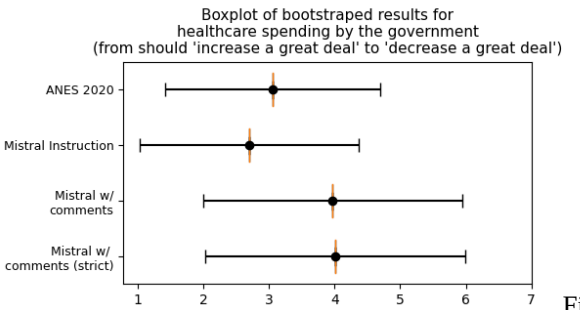


Figure 4: Boxplots of ANES (2020) and Mistral-7B-Instruct-v0.1 frequency distributions for social and Reddit sensing.